Data Wrangling Report

Date: 23/05/2020

1.1 **Introduction:**

The purpose of this project is to practice data wrangling skillsets, that we learned in Data Analyst Nanodegree. The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The steps of this project data wrangling, storing clean data and analyse and visualize those data, all are done in Jupyter Notebook.

1.2 **Data Wrangling:**

To wrangling the dataset we have to follows three steps -

- Gather Data
- Assess Data
- Clean Data

But for getting more clean and relevant large set of data, one can iterate through these processes again and again. Below we describe every step of the data wrangling process.

1.2.1 **Gather Data:**

For this project we gather data from the three different data sources. Those are follows -

- *Twitter archive file:* The twitter_archive_enhanced.csv is provided by Udacity and can be downloaded manually.

- *The tweet image predictions:* This flat file contains what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file image_predictions.tsv is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and theURL.
- *Twitter API:* Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Read this text file line by line into a pandas dataframe with tweet_id, retweet_count, favourite_count, followers count, friends count, source, retweeted_status and url.

1.2.2 **Assess Data:**

Assessment of gathered data is required for improving quality and tidiness of data, as most available data are messy and dirty and are not applicable for direct analysis.
During assessment one should take care of two features of data -

- *Quality* : In this case we check the completeness, validity, accuracy and consistency features of the data. And we have to modify the dataset to satisfy these features so that that can be applicable for analysis.
- *Tidiness:* In this case we look for the structure of the dataset. Untidy data known as messy data and this prevents easy analysis. Tidy data requirements are -
    1. Each variable forms a column.
    2. Each observation forms a row.
    3. Each type of observational unit forms a table.

    Here we also take care if there is data redundancy that can be avoided.

To satisfy the above described features of data, we assess our data in two ways -

- *Visual Assessment:* For visual assessment Jupyter Notebook and Spread Sheet are used. In jupyter notebook we print the three entire dataframe and in spread sheet we read the flat files.
- *Programmatic Assessment:* For programmatic assessment we use several Pandas methods (e.g. info, value_counts, sort_values, groupby, sample, duplicated etc.).

And we document the issues in two categories - quality issues and tidiness issues. During assessment our target is to make the dataset that contains the original rating and image along with other necessary information about tweets.

1.2.3 **Clean Data:**

Data cleaning process has three steps for individual issue, these are -

- *Define* - define the issue that we fix.
- *Code* - code to fix the issue.
- *Test* - test the issue is solved or not.

In this process we remove the issues those are documented in the assessment process.

The first and helpful step is to create a copy of the dataframe to perform the cleaning action, so that if we make mistakes or do something extremely wrong then we are able to create the same dataframe without a little effort.

In the copied version we do our cleaning process. Here we fix the issues we find in the data assessment process. During cleaning steps our target is to get a dataset that contains original posts with valid rating, image and all other necessary information regarding the post.

1.3 **Result:**

After wrangling the dataset we have a dataframe which has 1626 rows and 14 columns. This is a clean dataset on which we can do our required analysis and visualize the dataset. We store this clean data as a flat file "twitter_archive_master.csv" and in a table called "master" in a SQLite database called "twitter.db".

1.4 **Conclusion:**

Data wrangling is a core skill that whoever handles data should be familiar with. Data wrangling process makes a dataset analysis ready. Here for data wrangling we use the Python programming language and its various libraries, Jupyter Notebook and SpreadSheet. Also there are several tools for data wrangling, but python and its packages are more powerful than any other tools.